What Makes for Good Image Captions?

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper establishes a formal information-theoretic framework for image captioning, conceptualizing captions as *compressed linguistic representations* that selectively encode semantic units in images. Our framework posits that good image captions should balance three key aspects: informationally sufficient, minimally redundant, and readily comprehensible by humans. By formulating these aspects as quantitative measures with adjustable weights, our framework provides a flexible foundation for analyzing and optimizing image captioning systems across diverse task requirements. To demonstrate its applicability, we introduce the Pyramid of Captions (PoCa) method, which generates enriched captions by integrating local and global visual information. We present both theoretical proof that PoCa improves caption quality under certain assumptions, and empirical validation of its effectiveness across various image captioning models and datasets.

1 Introduction

Image captioning, the process of translating visual content into natural language descriptions, serving pivotal roles in real-world applications ranging from assisting visually impaired individuals [1–4] to facilitating content-based image retrieval [5–10]. Over the last decade, the field of image captioning has witnessed substantial progress, primarily driven by advancements in deep neural nets and the availability of large-scale high-quality image-text datasets.

Despite empirical advancements, several fundamental questions remain unanswered: *What makes for good image captions? Which properties should they possess, and how can we measure them?* Some existing models can generate captions closely resembling single-sentence human annotations [11], but these may not be adequate for use cases where more comprehensive coverage of fine-grained visual information is required. Conversely, recent Large Vision Language Models (LVLMs) [12, 13] have demonstrated the ability to generate multi-paragraph detailed image descriptions [14, 15, 12, 16]. Yet, longer captions can sometimes be less accurate, hallucinate content, or put excessive emphasis on irrelevant details while omitting important ones.

Recognizing the absence of a universal standard for ideal captions, this work aims to establish well-defined principles for image captioning that address varying task requirements. We introduce an information-theoretic framework based on semantic communication [17, 18] and the information bottleneck principle [19–21]. By leveraging this perspective, we formulate an objective function for image captioning that strikes a balance among three key criteria:

- **Information Sufficiency**: Ensuring comprehensive coverage of meaningful content, measured by the mutual information between the caption and task-relevant visual semantics.
- **Minimal Redundancy**: Optimizing the conciseness of the caption, quantified by the entropy of the generated caption.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

- **Human Comprehensibility**: Facilitating ease of understanding for human readers, assessed through the distributional distance between generated captions and natural language.

Our framework conceptualizes images and captions as observations of latent variables in a semantic space. This allows us to formulate image captioning as a communication process where semantic information is transmitted from the image to the caption, and measure the error at the semantic level. We then present formal quantitative measurements of the above three criteria and define the ultimate objective of image captioning as a weighted combination of them. Varying the weighting coefficients of these terms suits different preferences over image captions (*e.g.*, comprehensiveness, succinctness, readability), providing a rigorous foundation for analysis and evaluations.

Our framework provides a rigorous foundation for analyzing and advancing image captioning techniques. To demonstrate its practical applicability, we present the **Pyramid of Captions (PoCa)** method as an example application. PoCa employs a hierarchical approach to generate semantically rich captions by leveraging both local and global visual information. Utilizing our theoretical framework, we provide formal proof that each local-global aggregation operation in PoCa improves caption quality under certain assumptions. Empirical evaluations across various image captioning models and datasets corroborate our theoretical findings, showing that PoCa consistently yields more informative and semantically aligned captions while maintaining brevity and interpretability.

2 Proposed Framework

In this section, we provide a theoretical framework for image captioning as depicted in Figure 2. First, we formulate the task of image captioning by applying the concept of semantic units [17, 18]. We suppose that an image is an observation of a latent variable in a semantic space characterized by semantic units. An image captioning model will generate a caption for the image, and the caption can be mapped back to the latent semantic space and compared with the source semantic latent variable.

Based on this framework, we then introduce our proposed objectives inspired by the information bottleneck principle [19–21] for feature representation learning [21]. In our framework, we consider that the overall image captioning objective is composed of a *"information sufficiency"* term, a *"minimal redundancy"* term, and a *"human comprehensibility"* term.

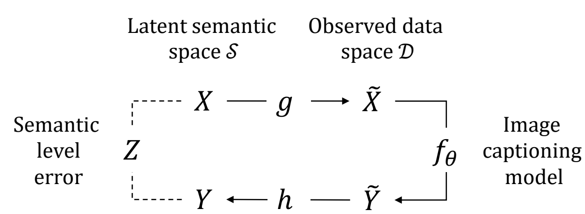


Figure 1: Overview of our formulation. Some latent variable X in a latent semantic space $\mathcal{S}$ generates image $\tilde{X}$ in data space $\mathcal{D}$. The image $\tilde{X}$ is then captioned by f_θ producing a caption $\tilde{Y}$ which can be mapped back to the original latent space as Y. The semantic-level error $Z = Y - X$ measures the difference between the source semantics X and received semantics Y.

2.1 Formulation

We assume that meaning arises from a set of independent and discrete units called semantic units in a semantic space, and images and captions are observations of some latent variables in this semantic space [17, 18]. Following [17, 18], we define a semantic unit and the semantic space as:

Definition 1 (Semantic Units and Semantic Space) *A semantic unit represents an atomic piece of information. The set of all possible semantic units is denoted by $\Omega = \{w_i\}_{i=1}^{n}$. A semantic space is an n-dimensional space $\mathcal{S} \in [0,1]^n$ where the value of the i-th dimension of represents the probability of the presence of the corresponding semantic unit $p(w_i)$.*

The Ω encompasses a wide range of semantic units, similar to how a vocabulary contains diverse words. Each point in $\mathcal{S}$ corresponds to a specific combination of semantic units, corresponding to different meanings. Adopting the semantic space and its probabilistic interpretation in [17, 22, 23], we can apply the classical information theory [24] to operate at the semantic level, and make information in images and captions to be comparable.

Let a random variable $X \in \mathcal{S}$ and $Y \in \mathcal{S}$ represent semantic information in an image and a caption, where $X_i = p(w_i)$ and $Y_i = p(w_i)$ represent the likelihood of a semantic unit w_i observed in an image $\tilde{X} \in \mathcal{D}_{\text{image}}$ or a caption $\tilde{Y} \in \mathcal{D}_{\text{caption}}$. These latent variables X and Y encode all the information within real images $\tilde{X} \in \mathcal{D}_{\text{image}}$ and textual captions $\tilde{Y} \in \mathcal{D}_{\text{caption}}$, both low-level and high-level semantics. Then, image captioning can be framed as follows:

Definition 2 (Image Captioning) *An image captioning model f parameterized by θ operates in the observed data spaces $f_\theta : \mathcal{D}_{\text{image}} \to \mathcal{D}_{\text{caption}}$, it translates an image into a caption, i.e., $\tilde{Y} = f_\theta(\tilde{X})$. $\tilde{X}$ is generated from some source latent variable $\tilde{X} = g(X)$ while $\tilde{Y}$ can be converted back to the latent semantic space $Y = h(\tilde{Y})$. Let $Z = Y - X \in [-1, 1]^n$ denote the error between the source and recovered semantics caused by parameters θ.*

Here, Z can be associated with which kind of error $\tilde{Y}$ has; negative Z indicates that the caption misses some contents of the image (undercoverage), and positive Z indicates that the caption includes something that is not in the image (hallucination).

2.2 Objectives

The image captioning process can be compared to a communication system [24] where information source X is converted to signal $\tilde{X}$ by a lossless source encoder g, transmitted through a noisy channel f_θ, and the received signal $\tilde{Y}$ is losslessly decoded by h, giving the final received information Y. From this communication system perspective, one might say that the optimal θ^* could just be the one that minimizes the error $||Z||$. However, this requirement is unrealistic as it would result in extremely long captions that losslessly encode both high-level semantic information and all the low-level irrelevant information.

Therefore, we apply the information bottleneck principle [19–21] to evaluate this system. Information bottleneck is a generalization of rate-distortion theory for lossy data compression, it requires a representation (which in our case is $\tilde{Y}$) to have maximal mutual information with some information T that is required fulfill the task requirements (*i.e.*, high information sufficiency), while having minimal mutual information with the input X (minimal redundancy). The desired minimal sufficient representation can be given as $\tilde{Y}^* = \arg\max I(\tilde{Y}; T) - \beta I(X; \tilde{Y})$ where β is a Lagrange multiplier. If the captioning model is deterministic (given X, it always produces the same $\tilde{Y}$) then we have $H(\tilde{Y}|X) = 0$. Since $I(X; \tilde{Y}) = H(\tilde{Y}) - H(\tilde{Y}|X)$, the minimal sufficient representation can be written as:

$$\tilde{Y}^* = \arg\max_{\tilde{Y}} I(\tilde{Y}; T) - \beta H(\tilde{Y}). \tag{1}$$

The second term will penalize the captioning model when it generates an over-length caption and the value of β controls the penalty strength. Combined with the first term, the objective requires the model to preserve as much useful information as possible for the task, while keeping the captions as succinct as possible.

Next, we give a formal definition of the information sufficiency objective of image captioning with importance of semantic units [17].

Definition 3 (Information Sufficiency) *For given X, let a latent variable $T \in \mathcal{S}$ represent the task-relevant information in X, and let an importance variable $A \in [0, 1]^n$ denote the importance scores of different semantic units. The A is derived from X by an underlying mapping, thus being dependent on X. The T is produced by a point-wise product between A and X, thus $T = A \odot X$. For generated image captions $\tilde{Y} = f_\theta(\tilde{X})$, the information sufficiency objective is:*

$$J_{\text{suf}}(\theta) = I(\tilde{Y}; A \odot X) \tag{2}$$

In the importance variable, $A_i = 1$ means the semantic units w_i are very important in the image, while $A_i = 0$ means w_i is irrelevant. It behaves similarly to the attention mechanism [25], which also produces a heatmap between zero to one according to the given input. Note that here the A is not binary, the continual nature of A gives a good property to $J_{\text{suf}}(\theta)$: when the "budget" is limited (as there are also other objectives to optimize), more semantic units with higher importance score will be retained while less important ones will be discarded.

Definition 4 (Minimal Redundancy) *The minimal redundancy objective encourages the image captioning model f_θ to eliminate irrelevant information, it is given by measuring the entropy of generated captions:*

$$J_{\min}(\theta) = -H(\tilde{Y}). \tag{3}$$

Combining $J_{\text{suf}}(\theta)$ and $J_{\min}(\theta)$ ensures the captions are **minimal sufficient representations** of images. However, there is no guarantee that the generated captions can be understood by humans. Therefore, we need to measure the human comprehensibility of the captions using a third objective, which is the distributional similarity between $Y\text{data}$ and natural language.

Definition 5 (Human Comprehensibility) *Let $P_{\tilde{Y}}$ denote the probabilistic distribution of model-generated captions over $\mathcal{D}_{\text{caption}}$, and let P_{lang} denote the distribution of human interpretable natural language. Given a certain statistical divergence measurement D, the human comprehensibility objective is:*

$$J_{\text{int}}(\theta) = -D(P_{\tilde{Y}} || P_{\text{lang}}). \tag{4}$$

The overall objective of image captioning is a weighted combination of information sufficiency, minimal redundancy, and human comprehensibility:

$$J(\theta) = J_{\text{suf}}(\theta) - \beta J_{\min}(\theta) - \gamma J_{\text{int}}(\theta), \tag{5}$$

where $\beta > 0$ and $\gamma > 0$ are weighting coefficients. Here, one of the factors that the coefficient β in Eq. 5 controls is the length of generated captions. If we prefer more detailed, comprehensive captions, we have smaller β and larger γ.

3 Example Application of the Framework

3.1 Method: The Pyramid of Captions

In this section, we introduce the Pyramid of Captions (PoCa) method, which showcases the applicability of our framework to image captionoing research. The key intuition behind the PoCa method is that we can have a more accurate and detailed caption by ensembling multiple captions.We propose to split an image into multiple local patches generate local captions for each patch, and fuse the captions to obtain a higher-quality caption for the global image.

Formally, let σ_{split} be a function operating in $\mathcal{D}_{\text{image}}$ that represents the splitting function, which splits an image into a set of local patches:

$$\sigma_{\text{split}}(\tilde{X}) = \left\{ \tilde{X}^{[j]} \right\}_{j=1}^{m}. \tag{6}$$

We apply an image captioning model f_θ to the local patches and obtain a set of local captions: $\{\tilde{Y}^{[j]} = f_\theta(\tilde{X}^{[j]})\}_{j=1}^{m}$, and also generate a caption for the global image: $\tilde{Y} = f_\theta(\tilde{X})$. The local and global information will be fused by a merging function σ_{merge} operating in $\mathcal{D}_{\text{caption}}$:

$$\tilde{Y}_{\text{merged}} = \sigma_{\text{merge}}\left(\tilde{Y}; \{\tilde{Y}^{[j]}\}_{j=1}^{m}\right). \tag{7}$$

We adopt text-only LLMs as the merging function σ_{merge}. Table 2 provides an example of merging using an LLM, where we instruct it to generate a merged caption that incorporates both local and global information.

As the PoCa method is hierarchical, we can extend it to be more layers. We can recursively split a patch into sub-patches and merge captions for each sub-patches to represent the patch.

3.2 PoCa Gives Better Captions (Provably)

In this section, we provide an analysis of under what condition a single local-global merging operation in PoCa can be guaranteed to improve the caption quality.

First, we assume that there is some function φ to quantify the error Z by X in a deterministic manner, and that function is concave. The deterministic assumption of φ simplifies the analysis by ignoring errors caused by factors other than the input semantics X, such as the randomness in sampling-based autoregressive generation. In other words, we assume that f_θ always generates the same caption and makes the same error for the same input. The concavity of φ implies that it generates a larger volume of error when $p(w_i)$ is far from zero and one. This means that the captioning model is more likely to make mistakes when there is high uncertainty about the presence or absence of a semantic unit in the image.

Assumption 1 (Uncertainty-aware content-dependent error) *The error Z produced by the image captioning model f_θ is dependent on the input semantics X. Therefore, it can be expressed as a deterministic function of X:*

$$||Z_i|| = \varphi(X_i), \tag{8}$$

where φ is a concave function and $1 \leq i \leq n$.

Next, we introduce our assumptions on the image splitting function σ_{split} and caption merging function σ_{merge}. These assumptions simplify the relationship between local and global semantics by assuming linear combinations. In practice, the relationship may be more complex and depend on factors such as the spatial arrangement of the local patches and the presence of objects spanning multiple patches.

Assumption 2 (Local-global relationship of image semantics) *The σ_{split} splits an image into local patches. The latent semantic variables corresponding to local patches satisfy the following relationship with global semantics:*

$$X = \sum_{j}^{m} \alpha_j X^{[j]}, \tag{9}$$

where the weights α_j satisfying $\sum_j^m \alpha_j = 1$.

Assumption 3 (Local-global aggregation of caption semantics) *The function σ_{merge} merges the global and global captions. The latent semantic variable corresponding to the merged caption is a weighted combination of the global and local semantics:*

$$Y_{\text{merged}} = \eta Y + (1-\eta) \sum_{j}^{m} \alpha_j Y^{[j]}, \tag{10}$$

where $\eta \in (0, 1)$ is a weighting coefficient.

We now present a theorem regarding $Z_{\text{merged}} = Y_{\text{merged}} - X$ (the proof can be found in the Appendix):

Theorem 1 (PoCa method reduces semantic error) *Under Assumptions 1-3, the PoCa method is guaranteed to have the smaller error Z_{merged} than Z, i.e.,*

$$||Z_{\text{merged}}|| \leq ||Z||. \tag{11}$$

Since A is non-negative, Theorem 1 implies non-decreasing information sufficiency; if merging does not increase redundancy or decrease interpretability, then the overall quality of caption becomes better. This also aligns with the findings in [26] that smaller-scale models combined can be as effective as a larger-scale model. Additionally, it is worth noting that our assumptions require linear combinations of semantics, while it may not hold in practice for images with more complex structures of semantics.

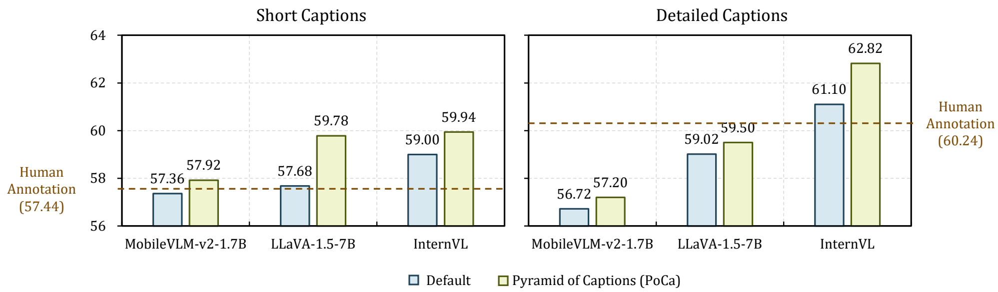


Figure 2: **Evaluation of Sufficiency**. VQA accuracy using captions generated by different image captioning models and the proposed PoCa method. The PoCa method consistently improves the VQA performance across all models and caption types.

3.3 PoCa Gives Better Captions (Empirically)

We conduct quantitative evaluations to study whether the PoCa method can improve the caption quality. We adopt the VQA-v2 [27] dataset, which is built upon the MS-COCO [11] dataset and contains multiple questions per image. The questions serve as a proxy for the importance score A in Definition 3, and the accuracy of the text-only LLM (LLaMA2-Chat-13B, prompt shown in Appendix Table 5) generated answers becomes an estimation of the information sufficiency term. See Appendix for more implementation details.

Figure 3.3 provides the evaluation results on 5,000 questions in the VQA-v2 validation split. As can be seen, our proposed PoCa method (green) consistently yields performance gains across all three examined LVLMs. The scale of improvement ranges from 0.48% (MobileVLM-v2 detailed captions) to 2.10% (LLaVA-1.5 short captions). Interestingly, we find that detailed captions do not necessarily correlate with better information coverage, as the detailed captions generated by MobileVLM-v2 underperform the single-sentence captions generated by InternVL. The comparison also shows that human annotations may not be optimal for certain scenarios, since several groups of LVLM-generated captions can yield higher VQA accuracy compared to that of human annotators.

It is crucial to evaluate whether the performance gain brought by PoCa is achieved by significantly sacrificing other objectives. In Table 3.3, we present the length statistics. We calculate the average number of words in default captions and PoCa captions and note their differences in the "$\pm\Delta$" column. The results show that PoCa does not exhibit a significant trend of either increasing or decreasing the length of captions. Among the six comparisons, PoCa compresses the length in four cases and extends the length in two cases. This empirically demonstrates that using LLMs as σ_{merge} in the PoCa model does not significantly violate the minimal redundancy objective $H(\tilde{Y}_{\text{merged}})$.

Table 1: **Evaluation of Redundancy**. Caption lengths statics between default and PoCa captions.

LVLM	**VQAv2** [27]			**Img2P** [28]		
	Default	**PoCA**	$\pm\Delta$	**Default**	**PoCA**	$\pm\Delta$
MobileVLM-v2-1.7B	54.1	78.2	+24.1	61.6	47.0	-14.6
LLaVA-1.5-7B	82.7	74.7	-8.0	93.2	133.4	+40.2
InternVL	158.3	93.4	-65.0	177.4	176.2	-1.2

4 Conclusion

Our work presents a novel information-theoretic framework that provides well-defined principles for image captioning covering information sufficiency, minimal redundancy, and human interpretability. By leveraging the theoretical framework, we propose Pyramid of Captions (PoCa), a novel image captioning approach that employs a hierarchical method to generate content-rich captions by exploiting the complementary nature of local and global visual cues. Through theoretical proofs and empirical evaluations, we demonstrate that PoCa consistently enhances the quality of image captions, making them more informative, semantically accurate, and contextually coherent while maintaining brevity and interoperability.

References

[1] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer, 2020.

[2] Chinmayi Rane, Amol Lashkare, Aarti Karande, and YS Rao. Image captioning based smart navigation system for visually impaired. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–5. IEEE, 2021.

[3] Yu Guo, Yue Chen, Yuanyan Xie, Xiaojuan Ban, and Mohammad S Obaidat. An offline assistance tool for visually impaired people based on image captioning. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 969–976. IEEE, 2022.

[4] Jothi Ganesan, Ahmad Taher Azar, Shrooq Alsenan, Nashwa Ahmad Kamal, Basit Qureshi, and Aboul Ella Hassanien. Deep learning reader for visually impaired. *Electronics*, 11(20):3335, 2022.

[5] Venkat N Gudivada and Vijay V Raghavan. Content based image retrieval systems. *Computer*, 28(9): 18–22, 1995.

[6] Rohini K Srihari. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9): 49–56, 1995.

[7] Ritendra Datta, Jia Li, and James Z Wang. Content-based image retrieval: approaches and trends of the new age. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 253–262, 2005.

[8] Ricardo da Silva Torres and Alexandre X Falcao. Content-based image retrieval: theory and applications. *RITA*, 13(2):161–185, 2006.

[9] Sahil Jain, Kiranmai Pulaparthi, and Chetan Fulara. Content based image retrieval. *Int. J. Adv. Eng. Glob. Technol*, 3:1251–1258, 2015.

[10] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. Recent developments of content-based image retrieval (cbir). *Neurocomputing*, 452:675–689, 2021.

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. URL `http://arxiv.org/abs/1504.00325`.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=w0H2xGHlkw`.

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *arXiv preprint arXiv:2312.08578*, 2023.

[15] Jaemin Cho, Yushi Hu, Jason Michael Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[16] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, Yushi Chen, Tong Lu, Jifeng Dai, and Yu Qiao. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=c2R7ajodcI`.

[17] Maxime Peyrard. A simple theoretical model of importance for summarization. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1059–1073. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1101. URL `https://doi.org/10.18653/v1/p19-1101`.

[18] Yixin Zhong. A theory of semantic information. *Proceedings*, 1(3), 2017. ISSN 2504-3900. doi: 10.3390/IS4SI-2017-04000. URL `https://www.mdpi.com/2504-3900/1/3/129`.

[19] Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *CoRR*, physics/0004057, 2000. URL `http://arxiv.org/abs/physics/0004057`.

[20] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017. URL `http://arxiv.org/abs/1703.00810`.

[21] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=-bdp_8Itjwp`.

[22] Yulin Shao, Qi Cao, and Deniz Gündüz. A theory of semantic communication. *CoRR*, abs/2212.01485, 2022. doi: 10.48550/ARXIV.2212.01485. URL `https://doi.org/10.48550/arXiv.2212.01485`.

[23] Kai Niu and Ping Zhang. A mathematical theory of semantic communication. *CoRR*, abs/2401.13387, 2024. doi: 10.48550/ARXIV.2401.13387. URL `https://doi.org/10.48550/arXiv.2401.13387`.

[24] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/J.1538-7305.1948.TB01338.X. URL `https://doi.org/10.1002/j.1538-7305.1948.tb01338.x`.

[25] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.0473`.

[26] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*, 2024.

[27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL `https://doi.org/10.1109/CVPR.2017.670`.

[28] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017.

[29] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, Jae Sung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, and Yejin Choi. Multimodal knowledge alignment with reinforcement learning. *CoRR*, abs/2205.12630, 2022. doi: 10.48550/ARXIV.2205.12630. URL `https://doi.org/10.48550/arXiv.2205.12630`.

[30] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4125–4134. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00425. URL `http://openaccess.thecvf.com/content_CVPR_2019/html/Feng_Unsupervised_Image_Captioning_CVPR_2019_paper.html`.

[31] Hao Liu, Wilson Yan, and Pieter Abbeel. Language quantized autoencoders: Towards unsupervised text-image alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/0df1738319f8c6e15b58cb16ea3cfa57-Abstract-Conference.html`.

[32] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A. Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, Kevin P. Murphy, Alexander G. Hauptmann, and Lu Jiang. SPAE: semantic pyramid autoencoder for multimodal generation with frozen llms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/a526cc8f6ffb74bedb6ff313e3fdb450-Abstract-Conference.html`.

[33] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[35] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[36] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[37] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=GUrhfTuf_3`.

[38] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[39] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=Ee277P3AYC`.

[40] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[42] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[44] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=WtmMyno9Tq2`.

[45] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[46] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023.

[47] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.

[48] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Enhanced visual instruction tuning for text-rich image understanding, 2024. URL `https://openreview.net/forum?id=tj4a1JY03u`.

[49] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: robot manipulation with multimodal prompts. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[50] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Gpt-4v (ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*, 2023.

[51] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=5KojubHBr8`.

[52] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 17773–17794. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/398ae57ed4fda79d0781c65c926d667b-Paper-Conference.pdf`.

[53] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.

[54] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

[55] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571, 2022.

[56] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, 2022.

[57] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14974–14983, 2023. doi: 10.1109/CVPR52729.2023.01438.

[58] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023.

[59] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023. doi: 10.48550/ARXIV.2310.03744. URL `https://doi.org/10.48550/arXiv.2310.03744`.

[60] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. Mobilevlm v2: Faster and stronger baseline for vision language model, 2024.

[61] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[62] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.595. URL `https://doi.org/10.18653/v1/2021.emnlp-main.595`.

[63] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics, 2005. URL `https://aclanthology.org/W05-0909/`.

[64] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P. Xing. Recurrent topic-transition GAN for visual paragraph generation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3382–3391. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.364. URL `https://doi.org/10.1109/ICCV.2017.364`.

[65] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Hierarchical scene graph encoder-decoder for image paragraph captioning. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 4181–4189. ACM, 2020. doi: 10.1145/3394171.3413859. URL `https://doi.org/10.1145/3394171.3413859`.

[66] Luke Melas-Kyriazi, Alexander M. Rush, and George Han. Training for diversity in image paragraph captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 757–761. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1084. URL `https://doi.org/10.18653/v1/d18-1084`.

[67] Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. Convolutional auto-encoding of sentence topics for image paragraph generation. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 940–946. ijcai.org, 2019. doi: 10.24963/IJCAI.2019/132. URL `https://doi.org/10.24963/ijcai.2019/132`.

[68] Jiwan Chung and Youngjae Yu. VLIS: unimodal language models guide multimodal language generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 700–721. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.46. URL `https://doi.org/10.18653/v1/2023.emnlp-main.46`.

[69] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/6fa4d985e7c434002fb6289ab9b2d654-Abstract-Conference.html`.

[70] Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17745–17753. AAAI Press, 2024. doi: 10.1609/AAAI.V38I16.29727. URL `https://doi.org/10.1609/aaai.v38i16.29727`.

Appendix

A Related Work

A.1 Image Captioning

Image captioning lies at the intersection of computer vision and natural language processing, requiring both accurate visual recognition and coherent language generation abilities. In Section 2 we formally defined the objectives for this task, but it is worth noting that current efforts do not explicitly optimize that objective. The primary challenge is the difficulty of back-propagation through the discrete textual space $\mathcal{D}_{\text{caption}}$, and efforts addressing this challenge involve adopting reinforcement learning [29, 30] or aligning continuous latent spaces with language spaces [31, 32]. However, these approaches suffer from training instability and less satisfactory language coherence.

Most current methods rely on a surrogate methodology, where they use human annotators f_{human} to write captions and train models to imitate those captions. The underlying assumption is that human-written captions optimize the objective $J(\text{human})$, which is achieved by providing instructions to crowd-sourced caption annotators. For example, the instructions for MS COCO Caption annotation [11] include "Describe all the important parts of the scene" and "Do not describe unimportant details", which are respectively connected to the information sufficiency term and minimal redundancy term in our objective.

An important trend in the image captioning field is the increasing focus on the comprehensiveness of image captions. As mentioned earlier, this represents a decreased length penalty (smaller weight β for the minimal redundancy term) and more emphasis on the information sufficiency term. In recent years, there has been an increasing number of high-quality detailed captioning datasets for this target, such as human-annotated image paragraph captioning [28], Densely Captioned Images (DCI) [14], and pseudo-labeled datasets, including LLaVA-Detailed-Captions [12], ShareGPT4V [33], AS-1B [16], etc. However, detailed caption annotation is much more expensive than previous single-sentence annotation, while automated caption labeling exhibits a high risk of hallucination.

A.2 Vision-Language Learning in the Era of Large Language Models

Various methods have been explored for enabling vision-language learning in LLMs. One line of work focusing on vision-language alignment during pretraining [34–43], allowing the model to jointly learn a shared vision-language latent space. The other line of work, improve the vision-language training efficiency by aligning the vision representation into the language space of LLMs by only training the visual encoder module or a vision-language projection matrix [44, 45, 12, 46–48]. These two lines of works enable vision-language alignment, enabling various joint vision and language modalities prompting methods such as robot manipulation prompting [49, 50] and multimodal in-context learning [51, 52].

Unlike the other two directions, another line of work exploits the reasoning and planning ability of LLMs allowing zero-shot multimodal vision-language inference by extracting the information from the vision modality into a textual description and performing inference through frozen LLMs [53–56]. Recent works in this direction showcase remarkable VQA ability through answer heuristics generation [57] and enabling object tagging and image editing through visual programming [58]. Inspired by this line of work, our work introduces a zero-shot hierarchical image captioning approach that relies on the reasoning ability of LLMs to aggregate information from local and global captions.

B Proof for Theorem 1

Theorem 1 (PoCa method reduces semantic error) *Under Assumptions 1-3, the PoCa method is guaranteed to have a smaller error $Z_{\text{merged}} = Y_{\text{merged}} - X$ than Z, i.e.,*

$$||Z_{\text{merged}}|| \leq ||Z||. \tag{12}$$

Proof 1 *First, we express the error of the i-th semantic unit in merged caption $Z_{\text{merged},i}$ as the difference between the merged caption semantics $Y_{\text{merged},i}$ and the source semantics X_i. Using*

Assumption 3 and 2, $Z_{\text{merged},i}$ can be expressed as:

$$
\begin{aligned}
Z_{\text{merged},i} &= Y_{\text{merged},i} - X_i && (13)\\
&= \eta Y_i + (1-\eta)\sum_j^m \alpha_j Y_i^{[j]} - X_i && (14)\\
&= \eta(X_i + Z_i) + (1-\eta)\sum_j^m \alpha_j (X_i^{[j]} + Z_i^{[j]}) - X_i && (15)\\
&= \eta Z_i + (1-\eta)\sum_j^m \alpha_j Z_i^{[j]}. && (16)
\end{aligned}
$$

Here, (15) is derived from (14) is based on the decomposition of the global and local caption semantics into there corresponding source semantics and errors, i.e., $Y_i = X_i + Z_i$ and $Y_i^{[j]} = X_i^{[j]} + Z_i^{[j]}$.

Next, we define $\Delta_{\text{PoC},i}$ as the gap between the norm of the global caption error and the norm of the merged caption error for the i-th semantic unit. Using the triangle inequality and Assumption 1, we derive a lower bound for $\Delta_{\text{PoC},i}$:

$$
\begin{aligned}
\Delta_{\text{PoC},i} &= ||Z_i|| - ||Z_{\text{merged},i}|| && (17)\\
&= ||Z_i|| - ||\eta Z_i + (1-\eta)\sum_j^m \alpha_j Z_i^{[j]}|| && (18)\\
&\geq ||Z_i|| - (\eta||Z_i|| + (1-\eta)\sum_j^m \alpha_j ||Z_i^{[j]}||) && (19)\\
&= (1-\eta)(||Z_i|| - \sum_j^m \alpha_j ||Z_i^{[j]}||) && (20)\\
&= (1-\eta)(\varphi(X_i) - \sum_j^m \alpha_j \varphi(X_i^{[j]})). && (21)
\end{aligned}
$$

To yield (19) from (18), we apply the triangle inequality, which states that for any two vectors a and b, $||a+b|| \leq ||a|| + ||b||$. Finally, we apply Assumption 1 to obtain (21), which states $||Z_i|| = \varphi(X_i)$ and $||Z_i^{[j]}|| = \varphi(X_i^{[j]})$.

By Assumption 1, φ is a concave function. Applying Jensen's inequality and Assumption 2, we have:

$$
\varphi(X_i) = \varphi(\sum_j^m \alpha_j X_i^{[j]}) \geq \sum_j^m \alpha_j \varphi(X_i^{[j]}). \tag{22}
$$

This inequality implies that the error of the global caption is always greater than or equal to the weighted average of the errors of the local captions. Intuitively, this means that the PoCa method, which combines information from both global and local captions, is expected to have a lower error than using only the global caption.

Combining this result with the lower bound for $\Delta_{\text{PoC},i}$ derived earlier, we can conclude that $\Delta_{\text{PoC},i}$ is non-negative for all i:

$$
\Delta_{\text{PoC},i} \geq (1-\eta)(\varphi(X_i) - \sum_j^m \alpha_j \varphi(X_i^{[j]})) \geq 0. \tag{23}
$$

The first inequality follows directly from the lower bound for $\Delta_{\text{PoC},i}$ derived earlier. The second inequality follows from the Jensen's inequality result, which states that $\varphi(X_i) \geq \sum_j^m \alpha_j \varphi(X_i^{[j]})$.

Since $1-\eta > 0$ (as $\eta \in (0,1)$ by Assumption 3), the product of $(1-\eta)$ and a non-negative term $(\varphi(X_i) - \sum_j^m \alpha_j \varphi(X_i^{[j]}))$ should also be non-negative, thus proving that $\Delta_{\mathrm{PoC},i} \geq 0$ for all i. Therefore, we have:

$$||Z_{\mathrm{merged}}|| \leq ||Z||. \tag{24}$$

This completes the proof, demonstrating that under the given assumptions, the PoCa method is guaranteed to reduce the semantic error compared to using only the global caption.

C Implementation Details

C.1 Image Captioning Models

We employ three groups of Large Vision Language Models (LVLMs) as the image captioning models: LLaVA-1.5 [59], MobileVLM v2 [60], and InternVL [61], among them:

- LLaVA-1.5 series is a popular LVLM with two variants: LLaVA-1.5-7B and LLaVA-1.5-13B, which adopt Vicuna-7B and Vicuna-13B as their Language Models (LLMs), respectively.
- MobileVLM v2 is a family of efficient LVLMs with smaller scales, and we utilize its MobileVLM-v2-1.7B and MobileVLM-v2-3B models.
- InternVL is one of the top-performing publicly available LVLMs. We use its InternVL-Chat-Chinese-V1-2-Plus model, which is based on the Yi-34B LLM and has a total of 40.1B parameters.

All inference is performed in FP16 precision on a single NVIDIA A800 GPU. We employ two types of prompts for short-form single-sentence image captioning and long-form detailed image captioning: `"Provide a one-sentence caption for the provided image"` and `"Describe this image in detail"`. All generation parameters are set to the default values provided by the source repository.

C.2 Caption Pyramids

For the caption merging function σ_{merge}, we adopt and compare a variety of Large Language Models (LLMs) as its implementation, including the Gemma family (2B and 7B versions), LLaMA2 family (7B chat and 13B chat), Qwen-1.5-7B Chat, Mistral 7B, and a mixture-of-expert model Mixtral 8x7B. The Mixtral 8x7B model has capabilities similar to ChatGPT-3.5 and is one of the top-performing open-source LLMs. All inference is performed in FP16 precision, except for the large Mixtral 8x7B model, for which we use 8-bit quantization to fit it into a single NVIDIA A800 GPU. The Mixtral 8x7B is used as the default LLM for caption merging.

We employ the prompt shown in Table 2 for caption merging, where the `"Assistant Generation Prefix"` is injected after the instructions to control the model output format. All generation parameters are set to the default values provided by the source repository. For splitting function σ_{merge}, we adopt the most straightforward implementation by splitting the input image into four equal-sized patches.

C.3 Human annotation baseline

In Fig. 3.3, the human annotation for short captions represents the accuracy of a single-sentence caption drawn from the MS-COCO annotation, while the human annotation for detailed captions refers to five MS-COCO caption annotations concatenated with the prefix `"The following are several captions of this image written by different people: "` added to the front.

Prompt for Merging Caption Pyramid

System Message:

Input:

- You will receive a **global caption** describing an image.
- Additionally, you will have access to **local captions** generated for specific patches within the image.
- Both global and local captions may contain noise or errors.

Task Objective:

- Your goal is to create a **merged global caption** that combines relevant information from both sources.
- The merged caption should be **no longer than the original ones**.
- You only give the merged caption as output, **without any additional information**.
- Do NOT give any explaination or notes on how you generate this caption.

Guidelines:

- **Combine Information**: Extract key details from both global and local captions.
- **Filter Noise**: Remove non-sense content, inaccuracies, and irrelevant information.
- **Prioritize Visual Details**: Highlight essential visual elements instead of feeling or atmosphere
- **Be Concise**: Use as few words as possible while maintaining coherence and clarity.
- **Ensure Coherence**: Arrange the merged information logically.

Remember, your output should be a high-quality caption that is concise, informative, and coherent!

User:

Global Caption: `{global caption}`
Top-left: `{top-left caption}`
Bottom-left: `{bottom-left caption}`
Top-right: `{top-right caption}`
Bottom-left: `{bottom-left caption}`

Assistant Generation Prefix:

Here's the merged caption:

Table 2: An Example implementation of the merging function σ_{merge} based on prompting text-only LLMs.

D Additional Experiments and Further Analysis

D.1 Image Paragraph Captioning

The image paragraph captioning dataset contains human-annotated single-paragraph descriptions for Visual Genome images. We use its testing split, which consists of 2,492 samples. We employ both the reference-free metric CLIPScore [62] and the reference-based metric METEOR [63] to evaluate the quality of the captions.

CLIPScore measures the similarity between the image and text features extracted by the CLIP model (we use the standard OpenAI pretrained `ViT-Base-32`). The underlying assumption is that CLIP encoders are capable of extracting semantic information and can represent the importance score A; thus, a higher CLIPScore correlates with higher information sufficiency.

The reference-based metric METEOR is widely adopted for evaluations in image captioning and natural language generation (*e.g.*, machine translation). It measures the word-level similarity between model-generated captions and human-generated captions. The underlying assumption is that human annotations optimize the information sufficiency objective, so if a model behaves similarly to human annotations, it achieves high information sufficiency.

The results are shown in Table 3, where we also list the performance of previous fully-supervised models and few-shot models. Once again, our PoCa method provides information sufficiency improvement according to both the reference-free metric CLIPScore and the reference-based metric

METEOR across all three families of LVLMs. These results further demonstrate the effectiveness of the PoCa method in enhancing the quality and informative content of the generated captions.

Table 3: Evaluation results on the image paragraph captioning dataset using CLIPScore and METEOR.

Image Captioning Model			CLIPScore	METEOR
Fully Supervised Models	Regions-Hierarchical [28]		-	15.95
	RTT-GAN [64]		-	17.12
	HSGED [65]		-	18.33
	SCST [66]		-	17.86
	CAE-LSTM [67]		-	18.82
Few-shot & Zero-shot Models	BLIP-2	3-shot	-	10.8
	OPT-IML	3-shot	-	9.5
	Naïve Ensemble	3-shot	-	9.8
	BLIP-2	VLIS [68]	-	14.6
	MobileVLM-v2-1.7B	Default	80.05	13.95
		PoCa	81.80	16.39
	MobileVLM-v2-3B	Default	79.02	8.99
		PoCa	81.34	13.28
	LLaVA-1.5-7B	Default	81.68	28.11
		PoCa	81.80	28.79
	LLaVA-1.5-13B	Default	82.16	28.44
		PoCa	82.47	28.97
	InternVL	Default	84.65	29.32
		PoCa	85.52	29.84

D.2 Caption Merging Strategies

In this section, we compare the effectiveness of different implementations of the merging function σ_{merge}. First, we compare various LLMs introduced in Section C.2. As shown in Table 4, compared to the global caption baseline, every LLM yields performance improvement, except for the smallest Gemma-2B-IT model. We also provide an ablation on prompting, where we replace the default prompt shown in Table 2 with a naive prompt of `"merge these captions"`. This ablation results in a slight decrease in accuracy and a significant increase in caption length, which further violates the minimal redundancy objective.

Additionally, we compare two parameter-free merging strategies based on simply concatenating local-only captions (representing $\eta = 0$ in Assumption 3) or local-global captions, with positional encoding as in the `"User"` field in Table 4. The results show that local captions alone cannot provide sufficient information, while adding the global caption brings significant improvement. However, these two concatenation-based methods generate excessively long captions, demonstrating the necessity of LLM-based information fusion and length compression.

Table 4: Comparison of different caption merging strategies on the VQA-v2 validation set.

Merging Function	Params	Accuracy	Length
Global Caption Baseline	0	57.68	50.75
Gemma-2B-IT	2B	57.44	107.14
Gemma-7B-IT	7B	58.74	178.79
Mistral 7B Instruct-v0.2	7B	58.92	136.12
LLaMA2-7B Chat	7B	58.64	199.34
LLaMA2-7B Chat (Naive Prompt)	7B	58.60	239.02
Qwen-1.5-7B-Chat	7B	58.64	130.93
LLaMA2 13B Chat	13B	59.06	154.67
Mixtral 8x7B Instruct-v0.1	46.7B	59.78	219.22
Local Captions Concatenation	0	55.66	265.63
Global Local Concatenation	0	59.12	337.38

D.3 Analysis of VQA-based Caption Evaluation

In the VQA-based evaluation, we adopt text-only LLMs for VQA inference with image captioning input to assess caption quality. This section provides a detailed analysis of this approach. Using

Prompt for LLM-based VQA Evaluation

System Message:
You will be given a caption of an image, and your task is to try to answer the question based on the caption. If the relevant information is not present in the caption, try your best to guess the answer. You shouldn't provide any rationale or explaination in your response, just give the answer only. The answer can be a number, a single word or a short phrase, plese make your response as short, simple and clear as possible.

User:
Image Caption: `{image caption}`
Question: `{question}`

Assistant Generation Prefix:
The most possible answer is:

Table 5: Prompt for LLM-based VQA Evaluation.

Table 6: Comparison of different LLMs for VQA-based caption evaluation.

LLM	Answer Length	Match Accuracy	NLI Accuracy
Gemma-2B-IT	33.50	5.20	55.44
Gemma-7B-IT	38.00	0.00	54.44
Mistral 7B Instruct-v0.2	28.90	2.30	63.30
LLaMA2 7B Chat	6.10	57.44	67.14
LLaMA2 7B Chat (No Caption)	4.30	41.34	44.76
Qwen-1.5-7B-Chat	7.90	56.72	69.06
LLaMA2 13B Chat	5.30	60.24	69.14
Mixtral 8x7B Instruct-v0.1	24.40	8.38	64.86
Ground Truth Answer	4.70	-	-

the instruction given in Table 5, we prompt different LLMs to generate answers and evaluate the accuracy based on exact matching and Natural Language Inference (NLI) based evaluation. The NLI evaluation classifies a pair of statements, `"The answer to this question is {ground truth}"` and `"The answer to this question is {generated answer}"`, into entailment, neutral, and contradiction, where entailment outputs are regarded as successful. Compared to exact matching, NLI evaluation measures the correctness of answers at a semantic level and can tolerate low-level differences.

As shown in Table 6, we find that different LLMs behave very differently in terms of answer length, and many LLMs fail to keep the answer succinct as instructed. Since the ground truth answers are mostly one word or a short phrase, this results in significantly reduced exact matching accuracy, while the actual semantic similarity is much higher, as measured by the NLI accuracy. We also observe an increasing trend in NLI accuracy when comparing different scales of LLMs, despite the largest Mixtral 8x7B Instruct-v0.1 producing lower NLI accuracy. We found that this outlier is caused by the over-conservative nature of the Mixtral 8x7B Instruct-v0.1 model, which frequently refuses to answer questions with responses such as "cannot determine" and "not sure". Finally, we add an ablation by instructing the LLM to guess the answer without caption input using the prompt: `"You will be given a question regarding an image, and your task is to try to infer the most possible answer"`. The resulting performance, noted as "LLaMA2 7B Chat (No Caption)", is much lower when measured by both exact matching and NLI accuracy.

Limitations

While the PoCa method has demonstrated effectiveness in improving image caption quality, there are several limitations that are worth discussing.

Assumptions on Image Semantics. The Assumption 2 made in this work could be sometimes strong and unrealistic, especially for the naive patch splitting function. The linear combination assumption may not hold well for images with more complex structures. This issue could be particularly problematic when objects or important semantic elements span across multiple local patches. In future work, employing more advanced splitting functions, object detection or semantic segmentation, could help alleviate this limitation and better capture the semantic structure of the image.

Assumptions on Caption Semantics. Similarly, the assumption about the local-global aggregation of caption semantics (Assumption 3) may not always be well satisfied by the LLM used for caption merging, particularly when the LLM is not sufficiently powerful. Weaker LLMs may struggle to effectively combine the local and global caption semantics in the desired manner. Further investigation into the impact of LLM choice on the fulfillment of this assumption would be valuable.

Depth of the Caption Pyramid. In the experiments, this work has demonstrated the benefits of a single level of local-global splitting and merging. However, the potential of deeper caption pyramids has not been fully explored. As the pyramid grows deeper, there could be a distribution shift for the input image patches, leading to more errors in the generated captions. Investigating the performance of merging functions for noisier captions is an important direction for future research.

VQA Evaluation. While the VQA-based evaluation provides a useful measure of caption quality in terms of information sufficiency, it has limitations. The questions used for evaluation may not comprehensively cover all of the important semantic units, resulting a sub-optimal estimation of the importance score A. In addition, due to resource constraints, we use a 5,000 question subset from the full VQAv2 dataset. To test its reliability, we run default caption generation with 5 models, together with human annotated caption, resulting in a total of $(5{+}1){\times}2{=}12$ data points combining short and long captioierns. The Pearson correlation coefficient between 5k subset accuracy and full dataset accuracy is 0.8519 – although already quite high, it still introduce some degree of noise for model performance evaluation.

Computational Efficiency. Our implementation of PoCa involves more inferences to generate captions and prompting LLM for fusing the local and global captions. These multiple inference steps and the use of large models can lead to increased computational costs. This computational overhead may be a concern, especially in resource-constrained environments or when processing a large number of images. One potential solution is to finetune an image captioning model on the captions generated by PoCa. By doing so, the knowledge captured by PoCa can be distilled into the finetuned model, allowing for a single inference pass during deployment, while still benefiting from the enhanced caption quality achieved by PoCa. Similar approach of knowledge distillation has been adopted in other literature, such as [33, 69, 70].

Broader Impact

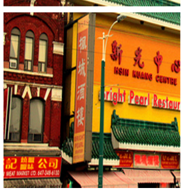
The proposed PoCa method has the potential to positively impact various applications, such as improving accessibility for visually impaired individuals and enhancing information retrieval. However, it is crucial to consider and address potential ethical implications, including biases inherited from LLMs, privacy concerns related to revealing sensitive information, and the environmental impact of training and deploying large-scale models. Responsible deployment, clear communication of capabilities and limitations, and ongoing efforts to mitigate biases and ensure fairness are essential for developing image captioning systems that are not only informative and accurate but also socially responsible and inclusive.

E PoCa Examples

We show some example of PoCa caption merging, where the images are sourced from the test split of the image paragraph captioning dataset. Captions are generated by LLaVA-1.5-7B model and merged by Mixtral-8x7B LLM. It shows that local captions successfully captured some errors that are hallucinated or omitted by the global caption, while PoCa successfully aggregated such information (marked by **green**) and recompile it into a human-interpretable caption.

Global Caption The image features a street corner with a street sign hanging above the street. The sign is written in both English and Chinese, indicating that the location is in an Asian neighborhood. The street sign is positioned above a restaurant, which is likely the "Asian Pearl Restaurant." There are several people in the scene, with some standing near the restaurant and others walking around the area. A car is parked on the street, and a traffic light is visible in the background. The overall atmosphere of the scene suggests a bustling urban environment with a mix of cultures.

Top-left: The image features a **large, ornate building with a green roof**. The building has a prominent green roof, which stands out against the sky. The sky is a pale yellow, creating a contrasting backdrop for the building. The building's architecture and the green roof make it an eye-catching structure in the scene.



Bottom-left: The image features a brightly colored building with a **yellow and red facade, likely a Chinese restaurant**. The building is adorned with a large sign that reads "Bright Pearl." The sign is positioned above the entrance, making it easily noticeable. In addition to the main building, there are two smaller buildings visible in the scene, one on the left side and the other on the right side. The overall atmosphere of the scene is vibrant and inviting.

Top-right: The image features a street sign with a blue background and white lettering, indicating the name of the street. The sign is attached to a green pole, which is located near a traffic light. The traffic light is positioned on the right side of the pole, and it is currently displaying a red light. The scene appears to be set in a city, with the street sign and traffic light being essential elements of urban infrastructure.

Bottom-right: The image features a **street sign** in a foreign language, likely Chinese, **warning people about the consequences of dumping.** The sign is placed on a pole, and it is positioned near a building. The **sign is written in both English and Chinese**, making it accessible to a wider audience. In the background, there is a car parked on the street, and a person can be seen standing nearby. The scene appears to be set in an urban environment, with the street sign serving as a reminder for people to properly dispose of their waste.

PoCa Merged Caption: At a bustling Asian neighborhood street corner, an **ornate building with a green roof** and a **yellow-red Chinese restaurant** named "Bright Pearl" stand out. A street sign in English and Chinese indicates the location, while a **nearby sign warns against dumping**. A car, traffic light, and people add to the urban atmosphere. The likely "Asian Pearl Restaurant" is positioned near the street sign.

Human Annotation: A picture of a Asian town. There are many signs in an Asian language. Te sky above the town is gray. There is a large green pole to the right with signs mounted to it. There are multiple story buildings along the streets. There is a restaurant with the word pearl in it. Therre are blue and white signs on the pole.

Global Caption The image features a group of four people dressed in old-fashioned clothing, standing together in a train station. They are posing for a picture in front of a train, which is located on the left side of the scene. The people are dressed in various styles, with some wearing ties and hats. There are two handbags visible in the scene, one near the center and the other towards the right side. The overall atmosphere of the image suggests a vintage or historical setting.

Top-left: The image features a man standing in front of a train, which is prominently displayed in the scene. The train is positioned towards the right side of the image, and the man is standing closer to the left side. The man appears to be posing for a picture in front of the train. In the background, there is a building with a clock on its side, adding to the overall atmosphere of the scene. **Additionally, there are two other people in the image, one located near the left edge and the other towards the right side**.

Bottom-left: The image features a train station with a train on the tracks. There are **several people in the scene, some of them standing near the train, while others are scattered** throughout the station. **A man is standing close to the train**, possibly waiting for it to depart. In addition to the people, there are a few handbags visible in the scene, with one near the center of the image and another closer to the right side. The presence of these handbags suggests that the people in the station might be travelers or commuters.

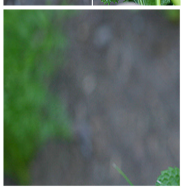
Top-right: The image features a group of three people dressed in **top hats** and Victorian-style clothing, standing together in a large room. They appear to be posing for a picture, possibly in a train station or a similar setting. The people are positioned in the center of the room, with one person on the left, another in the middle, and the third on the right side. The room has a few chairs placed around the area, with one chair located near the left side of the room, another in the middle, and a third chair on the right side. The overall atmosphere of the scene suggests a historical or themed event.

Bottom-right: The image features a man and a woman standing next to each other, both dressed in old-fashioned clothing. The woman is wearing a green dress, while the man is dressed in a suit with a tie. They appear to be posing for a picture, possibly in a historical setting. In the scene, there are also two handbags visible, one near the woman and the other closer to the man. The overall atmosphere of the image suggests a vintage or historical theme.

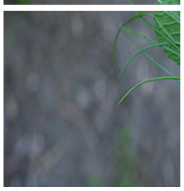
PoCa Merged Caption: A group of four people in old-fashioned clothing, including two in **top hats**, pose for a picture in a train station with a vintage atmosphere. A train is visible on the left, and two handbags are spotted near the center and right. **Two other people are scattered throughout the scene, with one man standing close to the train**.

Human Annotation: Four people wearing old-fashioned clothing stand posing in front of the camera. They are in a large building with a bright ceiling, and there is a black train behind them. There are two men on either side of two women. The man on the far left has a black suit, black shoes, and a black top hat. The woman to his right is wearing a white, pink, and green dress, with a white hat, and a broach pinned to her chest. She is holding a white satchel in her white-gloved hands. The woman next to her is also wearing white gloves and holding a white satchel. This woman has a green dress on, with lace on the trim. The man next to her is wearing a checkered vest underneath a dark long coat and dark pants. He has a black top hat.

Global Caption The image features a person holding a pair of scissors with a leafy green plant in their hand. The scissors are positioned in the middle of the plant, ready to cut it. The person is likely preparing to trim the plant or remove a leaf. The scene takes place outdoors, with the person standing in a garden or an outdoor area.



Top-left: The image features a close-up view of a patch of grass on a dirt ground. The grass appears to be green and healthy, with a few **small rocks** scattered around the area. The **dirt ground** is visible in the foreground, while the grassy patch extends into the background. The scene captures the essence of a natural environment, with the grass and rocks creating a sense of depth and texture.



Bottom-left: The image features a close-up of a green plant with a few leaves. The plant is situated in a dirt field, and its leaves are visible in the foreground. The plant appears to be growing in a **dirt-covered area**, possibly a garden or a natural setting. The focus of the image is on the plant and its leaves, creating a sense of depth and detail.

Top-right: The image features a close-up of a pair of scissors with a leafy green plant in the foreground. The scissors are positioned in the middle of the scene, with the leafy plant surrounding them. The plant appears to be a freshly cut herb, possibly **mint**, and is placed on a dirt surface. The scissors are likely being used to trim the plant or prepare it for use in a recipe.

Bottom-right: The image features a pair of green scissors with a green handle, placed on a pile of fresh herbs. The scissors are open, ready to be used for cutting the herbs. The herbs are scattered around the scissors, with some located closer to the scissors and others further away. The scene suggests that the person using the scissors is preparing to cut the herbs for cooking or other purposes.

PoCa Merged Caption: A person holds scissors with a leafy green plant, likely preparing to trim it in an outdoor setting. The scissors, situated in the middle of the plant, are positioned on a pile of fresh herbs. The plant, possibly a type of **mint**, appears **healthy and green**, **surrounded by small rocks and dirt**.

Human Annotation: There are a pair of scissors sitting on top of a plant. The handle on the scissors is colored green. The other part of the scissor is metal. The leaves of the plant or a nice healthy green color.

Global Caption The image depicts a group of people gathered around a dining table in a restaurant. They are enjoying a meal together, with several pizzas placed on the table. There are multiple cups and bottles scattered around the table, indicating that they are drinking beverages as well. The people are seated on various chairs placed around the table. Some of them are engaged in conversation, while others are focused on their meal. The atmosphere appears to be relaxed and social, as the group shares a meal and spends time together.

Top-left: The image depicts a group of people gathered in a room, possibly a restaurant or a bar. There are at least nine people visible in the scene, with some sitting at a dining table and others standing around the room. The room features a **spiral staircase**, which adds a unique architectural element to the space. A few chairs are placed around the dining table, and a cup can be seen on the table. There is also a bowl and a spoon placed nearby. The atmosphere appears to be lively and social, with people enjoying their time together.

Bottom-left: The image features a group of people sitting around a dining table, enjoying a meal together. There are at least five people visible in the scene, with some sitting closer to the table and others further away. The table is set with various food items, including a pizza, a bowl, and a sandwich. There are also multiple cups and a cell phone placed on the table. A bottle can be seen on the table as well. The dining table occupies a significant portion of the image, with the people sitting around it, creating a warm and inviting atmosphere.

Top-right: The image features a man with a long **beard** sitting at a dining table in a restaurant. He is looking at the camera, possibly posing for a picture. The man is wearing a blue shirt and appears to be enjoying his time at the restaurant. The dining area is furnished with multiple chairs and dining tables. There are two chairs visible in the scene, one near the man and another further away. Two dining tables can be seen, one in the foreground and another in the background. In the background, there is a TV mounted on the wall, likely providing entertainment for the restaurant's guests.

Bottom-right: The image features a person sitting at a dining table with a plate of pizza in front of them. The table is set with a glass of beer, a cup, and a pitcher of water. There are also two additional cups on the table. The person is holding a fork, ready to enjoy their meal. The dining table is surrounded by chairs, with one chair on the left side and another on the right side. The scene appears to be a casual dining experience, with the person enjoying their pizza and beer.

PoCa Merged Caption: A group of people, including a man with a **beard**, enjoy a meal in a restaurant with various pizzas, sandwiches, and beverages on a table. **Nine people** are gathered in a room with a **spiral staircase** and chairs around the dining table. The atmosphere is lively and social. (169 characters)

Human Annotation: three men are talking. they all have bears on their face. there is a blue shirt on the man. the shirt has flowers on it.